



Міністерство освіти і науки, молоді та спорту України

Національний технічний університет України

“Київський політехнічний інститут”

Фізико-Технічний інститут

### **Комп’ютерний практикум №1**

Криптографія

#### **Виконали:**

Студенти 3-го курсу ФТІ

групи ФБ-93

Тішков М.С. та Папуча Н.В.

#### **Перевірила:**

Селюх П.В.

## Мета роботи

Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

## Порядок виконання роботи

0. Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
1. Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку  $H_1$  та  $H_2$  за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення  $H_1$  та  $H_2$  на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення  $H_1$  та  $H_2$  на тому ж тексті, в якому вилучено всі пробіли.
2. Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела.

1. За допомогою написаною програмою отримаємо значення  $H_1$ ,  $H_2$ , таблиці з частотами монограмм та біграмм та надлишковість.

На вхід подаються два типа обробленого тексту: з пробілами та без

Такі результати були отримані:

**Для тексту без пробілів:**

Частота монограмм:

	Frequency
о	0,110783
е	0,085526
а	0,083096
т	0,06328
н	0,063122
и	0,062707
с	0,054542
л	0,049586
р	0,045731
в	0,044296
м	0,033788
к	0,033531
д	0,032297
у	0,028324
п	0,02596
я	0,020588
ь	0,020395
ы	0,018903
г	0,018788
б	0,01724
з	0,016917
ч	0,014943
й	0,010579
ж	0,010548
ш	0,007988
х	0,006472
ю	0,005959
э	0,004183
ц	0,003658
ф	0,003161
щ	0,003108

Частота біграмм без перетину (перші 10 найчастіші):

	Frequency
то	0,016521
ст	0,01293
на	0,010936
ал	0,010884
не	0,010644
но	0,010469
по	0,010413
ен	0,010203
ос	0,010126
го	0,009751

Частота біграмм з перетином (перші 10 найчастіші):

	Frequency
то	0,016641
ст	0,012903
на	0,010945
не	0,010761
ал	0,010751
по	0,010457
но	0,010381
ен	0,010228
ос	0,010173
ов	0,009843

Для тексту з пробілами:

Частота монограмм:

	Frequency
	0,161859
о	0,092852
е	0,071683
а	0,069646
т	0,053037
н	0,052905
и	0,052557
с	0,045714
л	0,04156
р	0,038329
в	0,037126
м	0,028319
к	0,028103
д	0,027069
у	0,02374
п	0,021759
я	0,017256
ь	0,017094
ы	0,015844
г	0,015747
б	0,014449
з	0,014179
ч	0,012525
й	0,008866
ж	0,008841
ш	0,006695
х	0,005424
ю	0,004995
э	0,003506
ц	0,003066
ф	0,002649
щ	0,002605

Частота біграмм без перетину (перші 10 найчастіші):

	Frequency
о_	0,020281
е_	0,018154
_в	0,017332
_с	0,01566
и_	0,015535
а_	0,015449
_п	0,015049
_н	0,013748
то	0,013664
я_	0,011048

Частота біграмм з перетином (перші 10 найчастіші):

	Frequency
о_	0,020318
е_	0,017958
_в	0,017329
_с	0,015704
и_	0,015545
а_	0,015415
_п	0,015152
_н	0,013718
то	0,013632
я_	0,011049

	Текст без пробілів	Текст з пробілами
H1	4.462	4.379
Redundant (H1)	9.917%	15.293%
H2	4.145204	3.972289
Redundant(H2)	16.329%	23.748%
H2 (with intersec.)	4.145589	3.972717
Redundant (H2 with intersec.)	16.321%	24.299%

2. За допомогою програми CoolPinkProgram оцінити значення  $H(10)$ ,  $H(20)$ ,  $H(30)$

[illegible]





$$R(\min) = 1 - 2.518 / 5 = 0.496$$

$$0.496 < R < 0.654$$

$$R(\text{cp}) = 0.575$$

$$1.613 < H(30) < 2.449$$

$$H(30)_{\text{cp}} = 2.031$$

*Надлишковість мови R для H(30):*

$$R(\max) = 1 - 1.613 / 5 = 0.677$$

$$R(\min) = 1 - 2.449 / 5 = 0.510$$

$$0.510 < R < 0.677$$

$$R(\text{cp}) = 0.594$$

$$R = 1 - \frac{H_{\infty}}{H_0}$$

$$H_0 = \log_2 32 = 5$$

**Середнє значення надлишковості:**

$$R = (0.486 + 0.575 + 0.594) / 3 = 0.552$$

**Висновок:** виконуючи дану лабораторну роботу ми набули практичних навичок щодо оцінки ентропії, у нас з'явилася справжнє уявлення про ентропію, про надлишковість мови.

При використанні власноруч написаної програми були отримані данні в яких можна помітити велику різницю між даними від двох різних видів обробки тексту. Дана різниця пов'язана із тим, що текст має дуже велику кількість пустих рядків із пробілами.

За допомогою програми CoolPinkProgram оцінити значення H (10), H (20) , H (30). Для кожного знайшли надлишковість, в наших експериментах R середній вийшов 0.552 (~55%). Але це природня надлишковість, як і всіх інших мовах.